A Satisfaction-based Model for Affect Recognition from Conversational Features in Spoken Dialog Systems

Syaheerah Lebai Lutfi , Fernando Fernández-Martínez , Juan Manuel Lucas-Cuesta , Lorena López-Lebón , Juan Manuel Montero

Abstract

Detecting user affect automatically during real-time conversation is the main challenge towards our greater aim of infusing social intelligence into a natural-language mixed-initiative High-Fidelity (Hi-Fi) audio control spoken dialog agent. In recent years, studies on affect detection from voice have moved on to using realistic, non-acted data, which is subtler. However, it is more challenging to perceive subtler emotions and this is demonstrated in tasks such as labelling and machine prediction. This paper attempts to address part of this challenge by considering the role of *user satisfaction ratings* and also *conversational/dialog features* in discriminating contentment and frustration, two types of emotions that are known to be prevalent within spoken human-computer interaction. However, given the laboratory constraints, users might be *positively biased* when rating the system, indirectly making the reliability of the satisfaction data questionable. Machine learning experiments were conducted on two datasets, users and annotators, which were then compared in order to assess the reliability of these datasets. Our results indicated that standard classifiers were significantly more successful in discriminating the abovementioned emotions and their intensities (reflected by user satisfaction ratings) from annotator data than

[*] Corresponding author

[**] Principal corresponding author

Email addresses: syaheerah@die.upm.es (Syaheerah Lebai Lutfi), ffm@tsc.uc3m.es (Fernando Fernández-Martínez), juanmak@die.upm.es (Juan Manuel Lucas-Cuesta), lorena.llebon@alumnos.upm.es (Lorena López-Lebón), juancho@die.upm.es (Juan Manuel Montero)

URL: http://www.syaheerah.com (Syaheerah Lebai Lutfi)

from user data. These results corroborated that: first, satisfaction data could be used directly as an alternative target variable to model affect, and that they could be predicted exclusively by dialog features. Second, these were only true when trying to predict the abovementioned emotions using annotator's data, suggesting that user bias does exist in a laboratory-led evaluation.

Keywords: Automatic affect detection, affective spoken dialog system, domestic environment, HiFi agent, social intelligence, dialog features, conversational cues, user bias, predicting user satisfaction.

1. Introduction

Automatic affect detection of users during real-time conversation is the key challenge towards our greater aim of infusing affect into a natural-language mixed-initiative HiFi-control spoken dialog agent (henceforth 'HiFi agent'). The HiFi agent is a proprietary system of GTH (details in Fernández-Martínez et al. [2010a]).[1] The HiFi agent controls and manages the HiFi audio system, and for end users, its functions equate a remote control (select a CD, track or radio channel, record music, change channels etc.), except that instead of clicking, the user interacts with the agent using voice.

Converting a non-affective system into an affect-savvy one involves a fundamental challenge of robust automatic detection of user affect[2]. Traditionally studies on affect detection from voice are based on the acoustic-prosodical features, mostly using posed expressions [e.g. Banse and Scherer, 1996; Barra-Chicote et al., 2006, 2007; Grichkovtsova et al., 2012; Lutfi et al., 2009a; Nicholson et al., 2000;

[1] We are in the process of integrating the HiFi control spoken system with a recently developed task-independent emotional model called NEMO (described in [Lutfi et al., 2010, 2009b] for the ability to detect affect and adapt to it by means of emotional synthesized speech. A couple of demos showing the platforms of different domains used to test this model here: `http://www.syaheerah.com/?page_id=789`. Previously, we tested NEMO with the HiFi agent (HiFi-NEMO), particularly on how user affect would influence the HiFi agent's response generation using emotional speech synthesis developed by Barra-Chicote et al. [2010], we used constant values of the dialog features to inform affect. The next step would be to fully automatize affect detection by using the best classification scheme, whose details are presented in this paper.

[2] The terms 'affect' and 'emotion' carry the same notions and are used interchangeably in this paper. The same applies to 'dialog features' and 'conversational features'

Oudeyer, 2003; Toivanen et al., 2004] and more recently, using real-life data or authentic affective speech [e.g., Barra-Chicote et al., 2009; Forbes-Riley and Litman, 2011a] and [see the state-of-art in these special issue editorials Devillers and Campbell, 2011; Schuller et al., 2012]. These studies were commonly carried out in domains such as learning, call-centers and games and entertainment. Very few aim at identifying emotions that influence interactions within a domestic environment. In this paper, we present the work done in affect recognition within this domain.

There are several issues identified when using data based on authentic affective speech when modelling detection of affect. We highlight these issues in the following subsection, along with the approaches and hypotheses that have been formulated to address them.

1.1. Current problems faced in affect recognition using voice

1.1.1. Challenges in identifying and labelling mild emotions

Studies using corpora of posed emotions usually deal with the automatic prediction of basic or full blown emotions that are easily collected, often successfully discriminated with accuracies better than chance, even across languages and cultures [Pell et al., 2009] and served as models for real-time affect prediction. In reality though, these prototypical full blown emotions do not really frequently emerge within real-life affairs [Batliner et al., 2011; Laukka et al., 2011], much less within human-machine interactions, and were almost absent in real-life databases [Batliner et al., 2011]. Milder or subtler emotions are more likely to occur within spontaneous activities in everyday life (frustration or irritation, hesitation, boredom, empathy, serenity to name a few). For example, in the HiFi agent's first evaluation, no full-blown emotions were elicited by users, only mild colourings of certain emotions. However, it is more challenging to perceive milder emotions, even more to classify them automatically by means of machine learning techniques because of the vague boundaries between emotions. Realistic databases also usually have limited categories of emotions [Laukka et al., 2011; Vogt and André, 2005], and therefore speech with mild emotional expressions are commonly clustered and reduced as a two or three-class problem within an automatic classification task (e.g., positive vs negative emotion). Emotion intensities are also usually ignored.

First, the subtlety of emotions poses challenges in emotion labeling and machine learning tasks. A common approach using machine learning to detect affect in voice is by training the classifier to detect the correct emotion label of a given input [see Callejas and López-Cózar, 2008b; Devillers et al., 2002; Vidrascu and Devillers, 2005]. Labelled emotion tags are usually gathered from two or more independent expert or non-expert listeners. It is known that this method brings its own issues, which are briefly discussed in Section 3.1 (see the first paragraph). Instead of using the more problematic emotion tags as a target class, why not use the user satisfaction judgment directly? User satisfaction is a standard measure of how a system meets or surpasses user expectation and could represent the user's feelings towards the system (see related work in Section 5.1). This question leads to our first hypothesis:

- satisfaction judgment can be used in replacement of labelled emotion tags to directly model affect.

Second, one way to better detect subtle emotions as those mentioned earlier is by looking at non-visual or non-vocal cues that involve user conversational behaviour, such as how long the user takes (in terms of turns) to successfully obtain a particular request, or how the same request is being repeated in different ways (same request, different words), in an attempt to get the machine to intrepret the message correctly. In this paper, we explore the possiblity of detecting affect using dialog features without combining it with any other types of features (vocal, facial etc.,). Thus we also hypothesized that:

- user emotional states can be predicted from *conversational features* on their own (i.e. without the need of using other features, such as acoustic, facial, and so on), through user *satisfaction ratings.*

1.1.2. Challenges in collecting unmasked, bias-free data

Next, to use satisfaction data to model affect, we must first assure that the data is reliable - that it is representative of the users' real feelings towards the system (bias-free). Collecting "unmasked" data though, in a laboratory setting, comes with its own challanges.

It is strongly acknowledged that the artificiality of a laboratory environment poses a huge challenge when collecting unmasked emotional data [Picard, 1999]. Real-world usage is very difficult to simulate in a laboratory setting due to lack of contextual information (e.g., users are given certain objectives or missions to fulfill when interacting with an Spoken Language Dialog System (SLDS), representation of actual physical environment etc.). Evaluators tend to adapt themselves to the less natural setting, adjust their tolerance levels and mask their feelings or opinions of the system that is being evaluated (either through vocal, facial or even self-reported satisfaction). This phenomenon is known as positive user bias, and is discussed further in Section 3.3. Thus data is usually collected using a sample of users that is less representative. Though user biases in laboratory evaluations are known phenomena, they have not been empirically tested, at least not in the area of Affective Computing. This paper also addresses this issue - given the laboratory constraints discussed above, we hypothesized that:

- users are *positively biased* when rating the system that is being evaluated and therefore the reliability of the satisfaction data when modeling affect in an SLDS is questionable.

1.2. The corpus used

In this study, we have used a corpus (described in Section 6.2) that contains audio-video recordings of user-HiFi agent interactions in a laboratory setting. It should be noted that the first non-affective evaluation was conducted (using the non-adaptive HiFi agent) with the intention of only measuring the agent's performance (i.e., ability to execute the actions requested by users) [Fernández-Martínez et al., 2008], without forseeing the integration of any social intelligence.

2. Affective states accompanying interactions with domestic spoken dialog agents

Based on the observations of the interactions in the videos from past evaluations of the spoken dialog HiFi agent, we were able to identify a set of emotions that frequently occurred during user-HiFi agent interaction. Typical emotions involved were contentment, frustration, confusion and boredom. These emotions are within the same family of some of the basic emotions proposed by Ekman and Friesen [1978] namely happiness, anger, surprise and sadness respectively, but in finer and less intense nuances. One other emotion of interest was self-frustration, in which users displayed discontentment towards themselves for erroneously addressing the system. We also added neutral to represent situations where there was no particular emotion of the aforementioned type present. This paper would however focus on discriminating affect between two classes: *contentment* and *frustration*, two types of emotions that are known to be prevalent within spoken HCI. These two categories of affect represent positive and negative user emotional state and their varying intensities (e.g., at the end of an interaction, a particular user might have felt intensely content with the system when the user gave a score of 5 or 'excellent' (on a 5-point scale), and rather frustrated when he or she gave a score of 3) [3]

3. Research Motivations

3.1. Modeling affect by predicting user satisfaction

User satisfaction has been used as a classic measure of user opinions on a computer system, including SLDSs. Somehow studies concerning affective SLDS do not treat the user's *opinion* as a reflection of his or her *affect*. A different approach is usually adopted to investigate user emotions while interacting with a SLDS, commonly involving a manual labeling task; independent judges listen to the users' utterances and then label them with several emotion categories on a turn-to-turn basis. Human listeners do not usually achieve high agreements on these emotion classifications [Callejas and López-Cózar, 2008a,b], even when using trained judges [D'Mello et al., 2008]. Cowie et al. [2010] pointed out that challenges in using emotion labels are not only limited to ensuring that the labels are correct, but also that the raters *agree* on those labels. It has also been reported that perceived and actual states can be rather divergent [Tcherkassof et al., 2007], and the same goes for perceived and self-reported states [Truong et al., 2012].

It is quite apparent that the user satisfaction rating has thus far been ignored as an important variable to model users' emotional states. Conversely, in this paper, we show that user satisfaction rating could be used to capture users' impressions, but it is limited to the number and categories of emotions (along positive/negative axis) as well as the types of tasks involved.

3.2. Automatic affect detection from conversational features

While many studies focus on numerous channels for affect detection, very few have explored dialog as a potential source [D'Mello et al., 2008]. User affect could be mined from conversational elements, which are always cheaper and are usually obtained with few or no computational overheads. By looking for emotional information beyond the mainstream visual (facial, gesture, posture) and vocal elements (acoustical or prosodical), such as those extracted from conversational elements, one could combine these

[3] depending on the model that was chosen - different models have different groupings of scores, elaborated later in Section 8.3.2. A score of 3, for example, may either represent a low-intensity frustration (category Three version 2) or slight contentment (category Three version 1)

two elements into a single decision framework to infer a more meaningful social phenomenon. Often many socially related traits, such as age, culture and personality are detectable from the way a speaker interacts, and are not directly picked up from the words that are spoken [Grothendieck et al., 2009].

Additionally, in affective computing, using tailor-made databases with domain-specific information to model emotion is the usual practice [Cowie et al., 2010]. However, the conversational features identified in our study are the standard context-independent features that are normally collected in any speech systems, reducing the constraint of being domain-specific and making it more flexible to be used for training the models of the same types of emotions in other domains, without having to redesign the dialog manager.

3.3. Positive user bias in laboratory-led evaluations

Criticism of laboratory-led SLDS concerns positive bias, or users being acquiescent. Acquiescence bias holds that respondents to a questionnaire have a tendency to show agreeable behaviour or positive connotations [Podsakoff et al., 2003] out of politeness [Reeves and Nass, 1996; Saris et al., 2005] - due to the belief that the researcher has a positive judgment of his or her own product and differing with this judgment would be impolite to the researcher, or simply because it takes less effort to just favor the system regardless of its performance than carefully weighing each optional level of good and bad scores. It is noted that user bias is quite common especially in laboratory settings compared to the field environment users [Dybkjær et al., 2004] who do not have any 'moral' or imposed obligations to give positive judgments.

There are also concerns with regards to the use of predefined scenarios, in which users were denied the freedom of selecting the tasks on their own as they would have done in a non-constricted environment [Callejas and López-Cózar, 2008c] and that they stress on task-completion [Ai et al., 2007]. These reasons might have caused them to ignore certain aspects of the interaction, such as ease of interaction (or 'comfort factor', termed by Möller [2005]) and report a biased satisfaction rating. In our case this could be true - the fact that users were actually requested to address a certain number of goals in a predefined scenario (a 'mission-based' situation) might have caused them to ignore the ease of interaction. When an individual is given certain criteria (e.g: "You should put on the HiFi system") he or she tends to focus only on meeting the criteria for ultimate success, regardless of the consequences. Thus users might only be concerned about *whether* they have achieved a particular goal, but not with *how* it is being achieved. As long as their goals were met, users were satisfied, leading them to rate the agent's overall performance highly.

In order to ensure that the user satisfaction data in our corpus was reliable, we compared them with the ratings from a group of independent raters (offline users). Since these "offline users" (e.g., raters) were free from the constraints discussed above that concerned the users, there is a chance that they might give more impartial ratings.

4. Expected Contributions

The main contributions in this study could be summarized in three key points:

- To show that satisfaction data could be used as an alternative target variable for affect modeling.
- To show empirically that conversational features, a non-conventional source, could be used as a single source to model user affect reliably by predicting satisfaction ratings.
- To show empirically that users are positively biased when rating the system in a laboratory-led environment and therefore the reliability of the satisfaction data when modelling affect is questionable. We have also suggested a solution to this problem, precisely by reusing the same data in order to produce a valid finding.

5. Related work

5.1. On using satisfaction judgement to model affect

Several articles have defined satisfaction as an emotional response toward an object [Bailey and Pearson, 1983; Doll and Torkzadeh, 1988; Locke, 1976]. Specifically, Bailey and Pearson, and Doll and Torkzadeh described user satisfaction as a positive or negative sum of feelings or attitudes, affecting a specific situation - in short, an *affective attitude*. Several studies have also established empirical relationships between user satisfaction rating and emotion. For example, Gelbrich [2009] revealed that anger had a significant inverse effect on user satisfaction rating (using 10-point scales), within HCI in a self-serving device domain (i.e., mobile phone). In a similar vein, Kernbach and Schutte [2005] showed that the user satisfaction rating increased when service providers adapted themselves appropriately to the users' emotions.

There have been a few attempts at modeling affect by predicting satisfaction. Engelbrecht et al. [2009] attempted to model satisfaction as a variable of affect, using dialog and linguistic features, within a restaurant information SLDS. They used different prediction and test models, in which the prediction model was trained with data collected from people from the campus, and then tested on interaction data collected from expert laboratory researchers. Their approach is quite different from ours because they used exactly the same interaction scenarios across all users (which means that problems within the interactions might be constrained to a certain condition only) and also asked the users to rate their satisfactions towards the system on a *turn* basis, while in our study, the rating is based on a session basis. This idea is questionable because users would not have formulated their opinions of the system at least during the first few turns. Moreover, if a user felt negatively towards a system at the beginning of the interaction, the user would tend to remain in that negative state throughout the whole interaction, as pointed out in Riccardi and Hakkani-Tür [2005]. This indicates that if users give ratings on a turn basis, low ratings might be acquired for all the turns that follow a particular problematic turn without them waiting till the end of the interaction in order to formulate a sensible opinion of the system. Though all predictions were above the baseline, they were not significant. Later, they attempted to model satisfaction using acoustic features, documented in Burkhardt et al. [2009]. The accuracy improvement above the baserate is only about 5%, but by using a subset of the data (to ensure a normally distributed data) the improvement was 14%, however it is not significant.

5.2. On automatic affect detection from conversational features

Many previous efforts have leveraged discourse cues along with acoustic-prosodic and linguistic (i.e., lexical such as dialog acts) information [Ai et al., 2006; Ang et al., 2002; Callejas and López-Cózar, 2008a; Forbes-Riley and Litman, 2011a,b; Lee and Narayanan, 2005; Liscombe et al., 2005; Litman and Forbes-Riley, 2006]. Most of this work has reported improvements in classification accuracy of less than 5% over acoustical and lexical features. Other efforts have included problematic turns as emotion markers. For example, Callejas and López-Cózar [2008a] proposed an algorithm to distinguish between doubt and boredom by examining the location of problematic turns. In particular, they calculated the proportion of problematic turns (subdialog within the same goal) to the whole dialog history, which they termed *accumulated width* for repeated problematic turns that have the same goal and *depth* for the number of turns previous to that particular problematic turn. Using dialog features, the recognition accuracy improved by 12.7% over chance.

To our knowledge, there has only been one study that reported the use of dialog features exhaustively to detect affect [D'Mello et al., 2008], also including dialog acts. The ideas in their work were inspired by Porayska-Pomsta et al. [2008], both in learning context. D'Mello et al. reported statistically significant classification accuracy above chance level. However, there are three main differences between the present study and theirs. First, they analyzed the emotions of learners on a *turn basis* thus excluding the overall context that evolved within the interaction that might have led to different emotional experiences, whilst our analysis took into consideration the overall context of interaction, that might provide a more rational explanation of the user's emotion evolution. Secondly, they addressed a larger set of emotions, but did not account for the *intensities* of those emotions. Thirdly, their emotion annotation task involved *trained* judges, who were Facial Action Coding System (FACS) certified - these judges had been trained in reading emotion by detecting facial action units according to FACS, developed by Ekman and Friesen [1978] and provided several categories of *emotion labels* that were related to learning whereas in our study, we do not need to use the more sophisticated methods to produce statistically significant classification improvements. However it should also be noted that their study was within a learning domain, thus encompassing different user goals, preferences or expectations.

6. Method

6.1. User and Annotator studies

The potential use of conversational features as feasible indicators for affect detection was researched through the examination of data from two types of study: first, the data from a *user study* [Fernández-Martínez et al., 2010b]. In this study, participants interacted with our HiFi agent and at the end of each dialog session, they assessed the overall HiFi agent's performance by giving Global Satisfaction Scores (henceforth 'GSS') based on a 5 point Likert scale. All the interactions were audiovisually recorded and information from these interactions were logged automatically on the basis of various features as presented in Section 7 and stored in the HiFi-AV2 corpus. The GSS was then associated with various dialog features, which were mined from the HiFi-AV2 log files. Specifically, the GSS was assigned to each dialog feature vector that was obtained from the log file of an individual dialog session. For example, if the participant gave a GSS of 4 (*good*) to the first dialog session, the dialog feature vector extracted for that first session was associated with 'good'. All features were normalized by the dialog length (the total number of counts of turns in the dialog), as satisfaction ratings were collected at the end of the interaction in every scenario.

Second, the data from the *annotator study.* This study was similar to the first one, but this time it involved several independent annotators rating the system *and* judging users' emotion in each individual session. These were done by watching video recordings of users interacting with the system from the first previously mentioned study and then providing: (a) GSS based on the annotators' perspectives of user ratings and (b) suitable emotion labels based on the individual user's verbal and non-verbal cues (described in Section 6.3). Similarly, a dialog feature vector extracted from the individual participant's log file was then associated with both a GSS and an emotion category. This study served two main purposes:

1. To collect GSS-labeled data from two groups of independent raters (*user*) and (*annotator*) and comparing the findings of both ratings. Annotators' rating of the GSS is based on the perspective of the user, in other words the annotators were asked to imagine themselves as users when rating the system. Thus we could consider that the comparisons of both databases as that of users *actual* ratings and *targeted* ratings (GSS provided by annotators). This way we would be able to determine which source is more reliable to model users' affect and in turn, the *adaptation* of the HiFi agent to their affect.
2. To collect affect-labeled data (using several pre-defined emotion categories). This task was not covered in the first study. Classifiers were trained using both affect-labelled data *and* GSS independently. This present paper however addresses the findings from the study involving classifications performed only on the GSS labelled data of both annotators and users.

6.2. Description of HiFi-AV2 corpus

The HiFi-AV2 corpus contains audio-visual recordings of 19 expert users (12 males, 7 females). 'Expert' in this sense indicates those with good technical background and who are familiar with spoken dialog systems in general, but have no previous experience interacting with this HiFi agent in particular. The users were not given any incentives to participate in the evaluation. Each user was involved in 10 interactions of predefined (basic and advanced) and free scenarios, totalling 190 interaction sessions (190 records). This way, the user is able to explore the functionalities of the HiFi agent guided by the number of goals he or she should address (e.g., to start the system, play a certain CD from a certain CD player etc). In the *basic* set of scenarios, users were strictly guided and only had to address a single task - e.g. *"You should try to stop the CD from playing"*. In the *advanced* set users were less guided, and given a more complex combinations of tasks - e.g. *"You should attempt to play a track from the CD in an increased volume"*, and in the *free* set users were not constrained, given no restriction but were told that the tasks should focus on the three main devices contained within the HiFi system - the CD player, tape player or radio channel. The total speech length is 115 minutes, with average 7 minutes per speaker. More detailed description of this corpus were given elsewhere [Fernández-Martínez et al., 2010b].

6.3. Annotator study: HiFi-AV2 human annotation procedure

Out of 19 users, we chose 10 users on a random basis ($N_{session} = 100 =$ 100 records) to downsize manual labeling efforts. The interaction videos of these 10 users were then distributed among 17 independent expert annotators. As the users, the annotators were technically sound and were highly familiar with spoken dialog systems. Most of them *also* have considerable experience interacting with the HiFi agent. Each annotator was assigned to two or three users. Since each user has 10 interaction sessions, each annotators then annotated between 20 and 30 sessions. Each speaker was assigned to a minimum of 4 annotators. Figure 1 shows a couple of sample screenshots extracted from the videos that are used for annotation. The recordings were not split into shorter scenes (e.g turns-based scenarios), as done by some researchers [Callejas and López-Cózar, 2008c; Engelbrecht et al., 2009; Shami and Verhelst, 2007] to have the annotators focused on one turn at a time. This is to avoid losing contextual information in which the whole dialogue took place, because ideally in an affective system, linguistic information is an integral part of emotion detection and generation [Batliner et al., 2011]. The annotators were given a set of full recordings (from the start until the end of an interaction) and they were free to label as many defined emotions (as stated in Section 2) detected throughout the whole interaction. They were also asked to provide a GSS of the system within the scales of 1-5 (between 1: Very poor to 5: Excellent), just as the users did in the previous evaluation. It is important to note that the annotators were neither familiar with the users nor were they given any information about users' ratings as to not influence their own ratings. Additionally, annotators were asked to perceive the politeness of the system by assigning scores of the same scales.

Figure 1: Sample screenshots from the video recordings that are used for annotation.

The annotators' ratings were averaged per session and consequently used as the final label (targeted rating) for that session. Decimals are rounded to the nearest whole number.

7. Metrics of mixed-initiative HiFi-AV2 spoken dialog

A twofold laboratory-controlled evaluation process aimed at assessing the system both objectively and subjectively was conducted in the past [Fernández-Martínez et al., 2008]. In the objective evaluation, metrics that measure the dialog features were automatically collected - a log file is maintained at the end of each dialog session that captures the measurements of each of these metrics described below in Table 1. Conversely, the subjective assessment involved the gathering of metrics directly from users such as user satisfaction or opinion of the system. The latter was collected through a questionnaire, after each interaction. Both the objective and subjective metrics used in the evaluation of the HiFi agent are mostly adopted within the PARADISE framework [Walker et al., 2000].

The GSS given after each interaction session indicated the interlocutors' overall opinion of the system, which could be a good reference of their feelings about the system. Hence, the outcome is GSS and the predictors are a combination of quality and efficiency dialog metrics (further explained in Section 7.1) collected from the objective evaluation. Table 1 presents a summary of the objective metrics used.

Table 1: Dialog efficiency and quality metrics

Feature Aspect	Metric	Acronym	Description
Efficiency	Turns Taken	TT	Number of turns needed to complete a scenario.
	Contextual Turns	Context_Turn	Number of turns taken where contextual information handling strategies are applied successfully.
	System Requests	Sys_Req	Number of turns taken where the system requests missing information from the user.
	Executed Action	Exec_Act	Number of turns required to accomplish a particular goal (execute a specific action).
Quality (null-efficiency)	Help Request	Help_Req	User interrupts the interaction to request for some help.
	Cancellation Request	Can_Req	User promptly quits current interaction and starts a new one.
	Silence Timeouts	Sil_TO	Timeout occurs after silent phase of a given duration.
	Recognition Timeout	Recog_TO	Timeout occurs when recognition timer expires. E.g.: When user speaks lengthy sentence, and violates the time limit.
	System Failures	Sys_Fail	Occurs when the system failed to receive IR commands.
	Repeat Speech Recognition	Rep_Recog	User repeats an utterance and system captures newly recognized words in the repeated utterance.
	Repeat Speech Understanding	Rep_Semantic	User repeats an utterance that has the same semantic content.
	Speech Recognition Rejection	ASR_Rej	Occurs when words in an utterance obtain lower confidence score than certain threshold.
	Non-Language Understanding Rejection	NLU_Rej	Occurs when the concepts in an utterance obtain a lower confidence score than a certain threshold, albeit good overall recognition score.
	Out-of-domain words	OOD	occurs when words uttered are meaningless in view of dialog goal (i.e., the system is not able to determine any word that influences the execution of an action).
	Dialog Time	DialTime	Time required (in seconds) to complete a dialog.

7.1. Dialog quality and efficiency metrics

Both the objective and subjective metrics for evaluating technical and usability facets of SLDS such as those recommended in the International Telecommunication Union [Recommendation P.862, 2001] or SASSI [Hone and Graham, 2000] questionnaires are widely adopted by the speech research community as a de-facto standard for assessing spoken dialog systems across various domains. These questionnaires recommend a long list of metrics that address several aspects of spoken systems such as quality, efficiency, likeability etc.

Among these aspects, we emphasize both the *quality* and *efficiency* (a subset of quality) aspects. Measurements related to dialog management are often targeted in quality evaluation because their functionality is akin to usability [Dybkjær et al., 2004]. At such, quality measurements are extracted from subjective judgments by human users [Hone and Graham, 2000; Möller et al., 2007], and the values are then correlated to those from the objective metrics to pinpoint the features that explain the subjective results significantly. In our experiment, quality-related aspects were mainly (but not entirely) captured in:

- *Speech Recognition (ASR_Rej)* and *Understanding Rejections (NLU_Rej)* - these occur when either the recognition or understanding confidence value falls below a predefined threshold respectively,
- and the *Out-of-Domain Turns (OOD)* - occurs when sentences uttered are meaningless in view of dialog goal (i.e., these words do not trigger any action from the system).

In this study, these metrics are identified to have an important influence on dialogue efficiency, since each of them involves dialogue turns that do not result in any performed actions (thus termed as *null efficiency* metrics). Our corpus shows that 15.2% of the dialog consists of null-efficiency turns.

These are input-related metrics that provide the most basic information for spoken dialog systems, and are concerned with the speed and length of the interaction on which other features are dependent. Therefore the perceived quality of spoken dialog systems depend heavily on these quality-related metrics and is reflected in the user questionnaires. Möller et al. [2007] reported similar findings where ASR, NLU and system requests show high correlations with quality-related items in the questionnaire such as “difficulty with operation”, “system helpfulness”, “interaction pleasantness” and so on.

Additionally, we narrowed down several important metrics within quality metrics that are more interesting in view of efficiency. These metrics measure the *number of actions* that are executed per turn, the kinds of measures that are more tangible in view of the overall dialog quality:

- *Contextual Turns* - turns that rely on the contextual information resources for implicitly inferred information (see example in Table 2). It is measured as the percentage of turns in which contextual information handling strategies are successful.
- *System Requests* - turns where the system requests for missing or deliberately omitted information from the user. Also measured in percentage.
- *Turn Efficiency* - number of actions that are executed per turn.

Table 2: Context recovery using dialog history.

Turn (U:user; S:system)	**Details**
...	
U: “Play track number two”	
S: “Track number two is now playing”	System provides feedback
U: “Play number three”	The user omits the “track” parameter info
S: “Playing track number three”	The value given by the user unfortunately could match both “track” or “disc” parameters. However, the system is able to determine the correct one between them by checking the dialogue history from more recent to older entries and retrieving the newest one.
U: “Five”	Again referring to the “track” parameter info
S: “Track number five selected”	Once again the system elicits the correct parameter
...	

8. Inferring the user's emotional state based on the GSS judgement: Results and Discussions

8.1. Correlation between GSS and emotion-labeled data

The first question we had was whether the GSS can be used as an alternative target variable for modeling affect. To address this question, we first checked the correlation between the GSS and the affect-labeled data (both by annotators). We established a statistically significant correlation between both - r=.29. p<.01, suggesting that indeed, GSS could be possibly used as an alternative target to directly model affect. Though this is the case, further studies are required to determine whether GSS could be used as a *reliable* alternative target.

8.2. Multiple regression analyses

To determine the strength of the dialog features in predicting the GSS, multiple regression tests were conducted on the datasets from both the studies mentioned in Section 6.1, that include: all 19 participants (190 records - UserFULL) and randomly selected 10 participants (100 records - UserSEL) in user study, and the same selected participants in the annotator's study (100 records - AnnotSEL), leading to three multiple regression models. It is a well established fact that multicolinearity poses a problem in multiple regression in such a way as to produce unstable regression models [D'Mello et al., 2008; Field, 2005]. Multicollinearity exists when there is a strong correlation between two or more predictors, causing difficulties in assessing the individual importance or unique variance of a predictor and ultimately resulting in an increased variance of the regression coefficients [Field, 2005]. Therefore a collinearity diagnosis was conducted, where we removed strongly correlated dialog features on the basis of the Variance Inflation Factor (VIF) value (features with VIF greater than 10 are removed). This approach however only reduced the number of features by 1, discarding *Dialog Time*. Additionally, *Help Request*, *System Failure* and *Silence Timeout* were removed because of zero occurrence, yielding 11 dialog features.

Additionally, we organized the features into the subsets as shown in Table 1 and carried out the analyses in three steps in order to distinguish the variability among feature types. It was not really required in this study, however this additional analysis would give us an insight as to how much information could be extracted from each subset. This information might be important in the future, for example, should need arise for feature selection. In step 1, *null-efficiency* features (DF_{NE}) were included. In step 2, *efficiency* features were added (DF_{NE+E}) followed by the rest of the features in step

3. Table 3 presents the regression outcome for each of the feature subsets for each dataset:

- DF_{NE}: Model with null-efficient dialog features only (ASR + OOD + NLU)
- DF_{NE+E}: Model with null-efficient and efficient dialog features (all of the above + Exec_Act + TT + Req_Turn + Context_Turn).
- DF_{ALL}: Model with all dialog features (DF_{NE} + DF_{NE+E} + the rest of the dialog features).

For each of the models, the values for these variables (top row of Table 3) are given:

- the degrees of freedom (df),
- squared adjusted multiple coefficient correlation between the given conversational features and the satisfaction(R^2_{adj})
- the change in the probability distribution (F_{change}), and,
- the change in the squared adjusted multiple coefficient correlation between the given conversational features and the satisfaction(ΔR^2_{adj})

Table 3: Mutiple regression models for GSS in all three datasets

Dataset	Model	df1,df2	R^2_{adj}	F_{change}	ΔR^2_{adj}
userALL	DF_{NE}	3,186	.009	1.581	.009
	DF_{NE+E}	4,182	.211	**12.676**	**.202**
	DF_{ALL}	6,176	.230	1.766	.019
userSEL	DF_{NE}	3,96	.021	1.699	.021
	DF_{NE+E}	4,92	.236	**7.760**	**.215**
	DF_{ALL}	4,88	.261	1.786	.025
annotSEL	DF_{NE}	3,96	.063	3.228	.063
	DF_{NE+E}	4,92	.394	**14.100**	**.331**
	DF_{ALL}	4,88	.405	1.424	.011
Mean R^2_{adj} and ΔR^2_{adj}	DF_{NE}		.031		.031
	DF_{NE+E}		.280		**.249**
	DF_{ALL}		.298		.018

DF_{NE}: Model with null-efficient dialog features only (ASR + OOD + NLU).
DF_{NE+E}: Model with null-efficient and efficient dialog features (all of the above + Exec_Act + TT + Req_Turn + Context_Turn).
DF_{ALL}: Model with all dialog features (above mentioned features and the rest of the dialog features).
Bolded values: Models that are statistically significant at p <0.05.

8.2.1. Discussion on the multiple regression results

When aggregated across all three models, dialog features explained a total of 29.8% of the predictable variance in GSS regression, with 28% of the variance being accounted in step 2 alone, as presented in Table 3. Specifically, efficiency dialog features (DF_E) were discovered to be statistically significant at p <0.05 for all models, explaining 25% of the total variance. This result clearly suggests that efficiency features might be useful to predict users' opinion of the system. For both the experiment involving users - all 19 and the 10 selected users, efficiency features accounted for 20.2% and 21.5% (see the last column in Table 3, second and fifth row) of the total variances of 23% and 26.1% respectively. The strongest model was obtained for the annotator dataset, with efficiency features alone explaining 33.1% of the total variance of 40.5% - which also shows that annotators generally depended *one and a half times* as much as the actual users on conversational features in deciding the GSS.

8.2.2. Significant predictors of GSS

A number of significant predictors (at p<.05) of the coefficients in the multiple regression models in Table 3 and their positive or negative relationships with GSS (Table 4) allow us to attest several generalizations regarding the relationship between these features and users' feelings of the agent, which were reflected by user opinions on the system.

Table 4: Significant predictors in the regression models for all three datasets

Dialog features	Datasets		
	userALL	userSEL	AnnotSEL
Cancel_Req	−		
Recog_TO		+	
TT	−	−	−
Sys_Req			−
Context_Turn	−	−	−

userALL: User dataset with all 190 records, userSEL: User dataset with 100 selected records, annotSEL: Annotator dataset with 100 selected records.
+/−: feature is positive or negative predictor in the multiple regression model at p<.05 level.

A large number of turns (TT) affects the participants' satisfaction in a negative way, regardless of whether they belong to the user or annotator group. This finding is expected, as the greater the number of turns, the less positive impression users are going to have about the system [e.g. Charfuelán et al., 2000; Möller et al., 2007], or likely to get bored [e.g. Callejas and López-Cózar, 2008a; D'Mello et al., 2008]. Heightened dissatisfaction is also related to increased Cancel_Req and Sys_Req.

GSS increased when Recog_TO increased with the users (userSEL). This result can be first explained in view of a common problem in speech processing, which is voice activity detection (VAD). VAD is done typically by using several energy-based thresholds and deadlines to detect the presence of a voice (i.e., to identify the beginning or ending of a user utterance). However, this technique tends to be over-sensitive to background noise and non-speech sounds from the speaker, resulting in the HiFi agent erroneously detecting this noise as the start of an utterance. The detection of these false starts might trigger some unnecessary recognition timeouts. However, the agent did not give any feedback regarding the timeout but instead reacted by immediately performing the required actions based on its understanding of the partial user utterances that were captured[4]. The users might have viewed this kind of prompt response as the agent being helpful and efficient. Secondly, this view was also supported by video recordings that revealed users blaming themselves (conveyed via various gestures or facial expressions) for not addressing the system correctly, including being uncertain when giving requests.

Surprisingly, contextual turns (Context_Turn) negatively correlate with GSS. At first glance this may seem counter-intuitive mainly because contextual information was intended to expedite task delivery by introducing a repair strategy to reduce the number of system requests [Fernández-Martínez et al., 2010a,b]. Specifically, any ambiguous situation would be handled in such way that the system would recover the missing information from the dialog context in an attempt to deliver the task.

Unfortunately, the implemented strategies did not measure or check the appropriateness of the retrieved information. As a result, the system became extremely contextual (50% of utterances involved contextual turns). Almost every turn involved contextual information, and the number of contextual turns grew in parallel with dialog length. Discourse information builds up as the dialog evolves - the more turns taken, presumably the better context the system has over a particular interaction. However, Table 4 shows a result that is contrary to our assumption. Therefore, we decided to carry out a subsequent analysis by distinguishing dialogs based on the quality of their corresponding context. For example, short dialogs were grouped as those with "poor" context and likewise, long ones were grouped as those with "good" context.

The analysis on both the userSEL and annotSEL datasets revealed that though in general Context_Turn is negatively correlated with GSS, smoother and shorter dialogs (less than 5 turns[5]) have weaker correlations compared to those with more than 5 turns (see Table 5, first row).

This could be explained in the light of addressing goals in mission-based scenarios. The defined goals are expected to be achieved in a certain number of turns. Thus, possible problems during the interaction are evident in longer dialogs; the longer the user takes to address a particular goal, the higher

[4] This part actually highlights the agent's capabilities for real-time accurate understanding of partial utterances.

[5] the threshold of 5 turns is based on the decreasing trendline of the GSS at approximately 5 user turns.

Table 5: Correlation between GSS and Context_Turn based on different feature conditions for both datasets.

Feature	**Condition**	**userSEL**	**annotSEL**
Length	$\geqslant$ 5 turns	-.42	-.42
	$<$ 5 turns	-.21	-.36
Recog_accuracy	$\geqslant$ .60% (Good speakers)	-.40***	.05
	$<$ .60% (Other speakers)	-.18	-.38**

,* statistically significant at p<.05 and p<.01 respectively.

number of contextual turns and in turn the lower his or her satisfaction.

Also, though the result is not statistically significant, it gave interestingly suggestive differences between users and annotators, on some questions relating to the perception of contextual information for users who had short dialogs with the system and those who had longer ones. Users and annotators share the same perception in view of contextuality the same way for long dialogs but this is not the case for shorter ones - possibly pointing towards positive bias.

We have also contemplated the view that *user skill* could influence the way users perceive contextual information. For example, skillfull and more co-operative speakers may cause the efforts to interpret user requests to be much more reduced compared to the rest, as the first usually address the system in a more appropriate, succinct manner than the latter. To affirm this view, we split "good" speakers from the rest of the speakers of the userSEL dataset based on the criterion of *actual recognition accuracy* (a feature that was not included for GSS prediction and only used for analysis purposes)- speaker group with actual recognition accuracy greater than .60% were considered good and vice versa. However we were surprised to discover that contextual turns significantly explain good speakers' GSS, and the GSS were much more lower with the occurrence of contextual turns as compared to the other speakers (see Table 5, second row). Contextual information was intended to expedite task delivery, however these results indicated otherwise suggesting that the system might in fact have *inappropriately* or ambiguously interpreted user requests and introduced unnecessary additional turns as an attempt to retrieve the correct context. Moreover, good speakers might have given more conservative ratings compared to the bad ones as they might have felt that they have handled the conversation well with the system and were less tolerable to problematic interactions with the system. Annotators, however, had the view that is much more in line with our expectation, in that their ratings reflected that contextual information is almost irrelevant for good speakers (r=.05). Of course annotators were unaware of users' experiences but these results indicated that they were able to distinguish good speakers from the rest. As mentioned previously, good speakers held clear and orderly conversations with the system thereby providing the necessary information in every turn, decreasing the agent's effort in view of discourse strategy. Annotators also shared our view as regards "other speakers" where there was a substantial negative relationship between contextual turns and GSS - more turns taken in order for the system to grasp sufficient contextual information and thereupon expanding the dialogs. Indeed, the result was consistent when double-checked using the TT feature for other speakers; longer turns cause significantly low GSS (see Table 6).

Table 6: Correlation between GSS and TT based on different user factors for both datasets.

Feature	**userSEL**	**annotSEL**
Good speakers	-.21	-.12
Other speakers	-.28*	-.59***

*,*** statistically significant at p<.1 and p<.01 respectively.

Contrary to our assumption, the relationship between contextual turns and users' affect can turn out to be either negative or not correlated at all. A future step would also be to conduct a more exhaustive analysis such that grouping Context_Turn into variations of *appropriate* and *inappropriate* as done in the study of Danieli and Gerbino [1995]. In general though, the relationships between GSS and the aforementioned features are rather intuitive and follow the same numerical directions as expected.

8.3. GSS classification from conversation features

Real time automatic detection of emotion is vital to any affect-sensitive system [D'Mello et al., 2008; Picard, 1999]. Hence to detect affect via GSS based on dialog features, we applied standard classification techniques in which several classifier schemes were used with the intention of comparing the performance of the various classification techniques, in order to determine which technique(s) yield the best performance. The Waikato Environment and Knowledge Analysis (WEKA) [Witten and Frank, 2005] was used for these purposes. One or more classification algorithms were chosen from different categories including rule-based classifiers (ZeroR as baserate and OneR), functions (SimpleLogistic, SMO), meta classification schemes (Multischeme, MultiBoost, AdaBoost) and trees (J48). A 10-fold cross validation technique was used for all the classification tasks.

8.3.1. Positive bias in assessing HiFi SDS

The results from the experiment with the UserFULL dataset revealed no statistically significant result - at best, only 5 percent improvement from the baseline to OneR, revealing that GSS could not be predicted from the dialog features. This suggests the question of whether the users were rating the system randomly or were just being positively biased. Upon closer inspection of the data, we found that there were too few cases for point 1 (very poor) and point 2 (poor) categories, and majority cases turn out to have 4 (good) or 5 (excellent) points. This ceiling effect in reporting the GSS suggested that users might have been acquiescent when assessing the HiFi agent. In light of this discovery, we studied the correlation between GSS and the *actual recognition accuracy*, to confirm that the scores were biased. Weak correlation between the users' GSS and the actual recognition accuracy (r=.15) explained that users rated the system more favourably and were less critical towards the agent. Conversely , on the other hand, the *annotators* depended on this criterion significantly (r=.36, p<.01) to do the same.

The relationships above indeed confirmed that users have been undoubtedly biased or acquiescent. As for the annotators, since they were not provided with any predefined scenarios, they gave more impartial ratings.

8.3.2. Data redistribution

The reduced sample size for the userSEL and annotSEL datasets (from N=190 to N=100) resulted in unbalanced data - the datasets consist of some samples with less than two cases of the same outcome (e.g: only one single case of point 1 score for *very poor*). Thus to obtain a more uniform distribution, samples with similar outcomes were grouped together, and this was repeated five times to satisfy all combinations of classification problems as shown in Table 7. This way we were also able to determine which clusters obtained optimized classifications. To better view the skewed distribution of these labels, the distribution of each grouping over the entire set is shown.

Table 7: Datasets re-clustered according to similarity of score points into all possible combinations of classes

	Label				
Category	very poor	poor	satisfactory	good	excellent
Five (original class)	1	2	3	4	5
% distribution (U, A)	2, 1	5, 16	22, 36	33, 29	38, 18
Four	-	1,2	3	4	5
% distribution (U, A)		7, 17	22, 36	33, 29	38, 18
Three (version 1)	-	1,2	3	4,5	-
% distribution (U, A)		7, 17	22, 36	71, 49	
Three (version 2)	-	1,2,3	-	4	5
% distribution (U, A)		29, 53		33, 29	38, 18
Two (version 1)	-	1,2,3	-	4,5	-
% distribution (U, A)		29, 53		71, 47	
Two (version 2)	-	1,2	-	3,4,5	-
% distribution (U, A)		7, 17		93, 83	

All eight classifiers were evaluated on the UserSEL and AnnotSEL datasets across six categories. Table 8 presents the statistically significant improvements in classification results over baserate in percentage accuracy. The following section reports the effects from the analysis carried out using a factorial two-way independent analysis of variance (ANOVA).

.

Table 8: Comparisons of significant improvements of classifications in label accuracy (above) and F1-score (below) in detecting GSS from conversational features for both user and annotator datasets

Category	Classifiers																	
	Base rate		OneR		SiLog		SMO		Ord		MulS		MulB		AdaB		J48	
	U	A	U	A	U	A	U	A	U	A	U	A	U	A	U	A	U	A
Five	38.0	36.0	-	-	-	49.3	-	44.6	-	51.3	-	-	-	-	-	-	-	-
Four	38.0	36.0	-	-	-	53.1	-	43.4	-	52.0	-	-	-	-	-	-	-	-
Three (version 1)	71.0	47.0	-	-	-	64.0	-	61.1	-	62.5	-	-	-	-	-	-	-	59.0
Three (version 2)	38.0	53.0	-	-	-	-	50.7	-	-	-	-	-	-	-	-	-	-	-
Two (version 1)	71.0	53.0		61.3	-	75.0	-	74.4		69.4	-	-	-	71.2	-	74.4	-	69.4
Two (version 2)	93.0	83.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Five	0.21	0.19	0.43	0.39	0.36	0.46	0.40	0.35	-	0.47	-	-	0.28	0.29	0.28	0.29	-	0.39
Four	0.21	0.19	0.42	0.40	0.35	0.50	0.40	0.34	-	0.48	-	-	0.30	0.29	0.30	0.29	-	0.40
Three (version 1)	0.59	0.30	-	0.52	-	0.62	-	0.56	-	0.60	-	-	-	0.48	-	0.48	-	0.57
Three (version 2)	0.21	0.37	0.40	0.58	0.45	0.58	0.48	0.50	0.31	0.61	-	-	0.32	0.45	0.32	0.46	-	0.53
Two (version 1)	0.59	0.37	-	0.59	0.70	0.74	-	0.74	-	0.69	-	-	-	0.71	-	0.74	-	0.69
Two (version 2)	0.90	0.75	-	-	-	0.82	-	-	-	0.82	-	-	-	-	-	-	-	0.82

OneR= Rules.OneR, SiLog= Functions.SimpleLogistics, SMO= Functions.SMO, Ord= Meta.Ordinal, MulS= Meta.MultiScheme, MulB= Meta.MultiBoost, AdaB=Meta.Adaboost, J48= Trees.J48

U= Experiment using UserSEL data, A= Experiment using AnnotSEL data

Only results that are statistically significant at p<.05 are shown to facilitate readability

A three-factor independent ANOVA was performed in order to evaluate the performance of the classifiers in classifying GSS from the dialog features, and to investigate the main effects between each of the three factors and the improvement of classification accuracy above the baserate, as well as how these factors interact among each other. The first factor *subject type*, consists of two levels, user and annotator, second factor *classifier*, involved nine levels: all the various classifiers mentioned in Section 8.3. The last factor was the *class category*, composing six levels: the data regrouped into the categories presented in Table 7. The results are presented in Figure 2.

The results showed that there was no significant interaction between categories and classifiers.

8.3.3. Comparisons across subject types

ANOVA results revealed that there was a significant effect of the subject type (see Table 8, figure 2(a)), on the classification improvement (in % accuracy), $F(1,40)=83.07$, p<.001, partial $\eta^2=.68$. As indicated in figure 2(c), at least three classifiers that were evaluated on the annotator data showed significant improvement over the baserate in each category, with the exception to categories two and three (both version 2). On the other hand, the classifiers evaluated on user data mostly revealed worse results than baserate with exception of SMO, which improved significantly over baserate for category three (version 2) (see figure 2(c)). This indicates that most classifiers were able to predict the GSS from dialog features based on annotator data, suggesting that the annotators were more impartial when judging the HiFi agent.

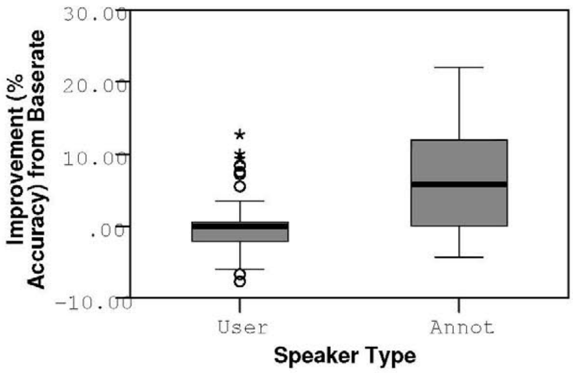
8.3.4. Comparisons across classifiers

There was a significant effect of the classifiers on the improvement of classification accuracy over the baserate, $F(8,40)=7.82$, p<.001, partial $\eta^2=.61$. Bonferonni *post hoc* test revealed that Simple Logistics classifier yielded the best performance ($M_{siLog}=7.72$, SD=8.04) and is significantly better (p<.05) than the meta and tree based classifiers (see Figure 2(b)).

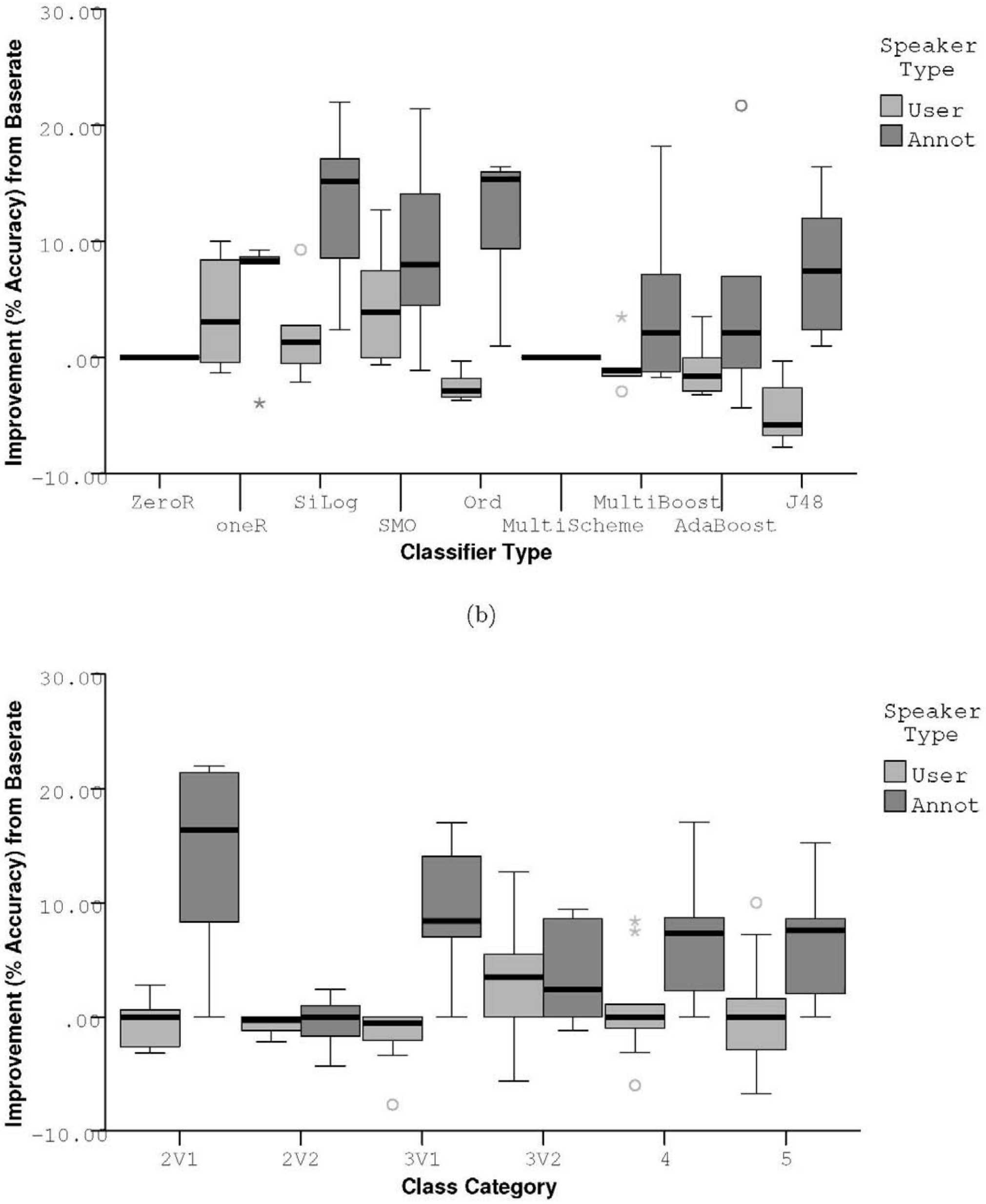
8.3.5. Comparisons across categories

The ANOVA results indicated that there was a significant main effect of the categories (various groupings) on the improvement of classification accuracy over the baserate, $F(5,40)=7.52$, p<.001, partial $\eta^2=.48$. Bonferonni *post hoc* pointed that the classifiers performed best when discriminating two classes (2V1), in which points 1,2 and 3 are collectively tagged as *poor* and point 4 and 5 as *good*, and was significantly higher than only category 2V2

(M_{cat2V1}=6.58, SD=9.7). However, when point 3 was tagged as *good* in the other version of the two-class problem (2V2), the result was contrary (see Figure 2(c)) - the classifers' performances were significantly worse than the rest of the categories (M_{cat2V2}=-.69, SD=1.63), suggesting that point 3 is a better representation of *poor* rather than *good*. In other words, when participants gave a GSS of point 3, they probably were mildly frustrated with the system, rather than mildly contented. Category 4 showed the next best improvement rate (M_{cat4}= 3.72, SD=6.27). Considering that the difference in the mean in the improvement rate between category 2V1 and 4 are not significant, this four-class classification model was chosen as the model for real-time affect detection for the affect-adaptable version of the HiFi agent (HiFi-NEMO) instead of the two-class (2V1) model, because classifiers are able to discriminate more varieties of classes. As such, this would also allow HiFi-NEMO to have a richer response generation model (to respond to requests with appropriate prosody, emotion intensity and content depending on the different perceived user affect).



(a)



(b)

(c)

Figure 2: Improvement accuracy in percentage by (a): subject type, (b): classifier type and (c): class category: 5: Five, 4: Four, 3V1: Three (version 1), 3V2: Three (version 2), 2V1: Two (version 1), 2V2: Two (version 2)

.

9. Conclusions and Current directions

This section concludes the findings contributed in this paper and briefly discusses the current directions of this work.

- *Modeling satisfaction as a variable of affect.* Our first main contribution is to show that GSS could be used as an alternative target variable for affect modeling, however limited to the number and types of emotions, and also the tasks involved within a particular domain. Several findings, including ours, revealed that satisfaction rating reflects user affect. We showed that there was a significant correlation between satisfaction and affect labeled data, although further study is required to determine whether GSS could be used as *reliable* predictors in place of emotion labels. In this study, ratings were given at the end of the interaction and not on the basis of turns, thus users and annotators have presumably captured a broader scope of contextual information that evolves over a series of turns leading them to experience certain emotions, mostly frustration or contentment. Most studies on affect-adaptive SLDS use the service of human raters to gather affect-labeled data, but often possess low interrater agreements. Following this, D'Mello et al. [2008] pointed out an interesting question; that if *humans* (especially trained raters) have a low consensus in classifying the emotions of others, how reliable can emotion classifications by machines be? Thus a future direction would be to compare emotion classifications that are evaluated on emotion-labeled data and those evaluated on satisfaction-labeled data (current results).

- *Using conversational features to model affect.* Our second main contribution in this paper is to show empirically that conversational features, a non-conventional source, could be used as a single source to model user affect reliably by predicting satisfaction ratings in HCI within a limited-task domestic domain. The conversational features were used as affect predictors and the GSS was the target. For this task we used an annotation method that is less sophisticated (such as the use of untrained judges to rate satisfaction instead of rating emotions) and smaller array of features for classification tasks. Nevertheless, emotion classification improvements achieved statistically significant results over baserate.

- *Bias detection.* Lastly, we showed that users are *positively biased* when rating the system and therefore the reliability of the satisfaction data when modeling affect in an SLDS is questionable. We empirically detected user bias within a laboratory-led evaluation. Whilst we demonstrated that in general, conversational features could predict frustration and contentment (and the intensities of these emotions) from satisfaction ratings, predicting them using data obtained directly from *users* were not possible. We found that users were inclined to inflate the agent's performance by evaluating the system favourably regardless of its actual performance, and thus 'masked' their satisfactions. It is a known fact that it is almost impossible to totally simulate a real world environment in a laboratory, and therefore laboratory data on emotions often cannot be generalized throughout the population. While we are not claiming external validity, we argue that the data could be reused in order to produce a valid finding. We did this by asking annotators to rate satisfactions as imaginary users. Classifications were evaluated on both these *actual* and *target* datasets. The results revealed that satisfaction from the latter were significantly predictable, but not from the former, suggesting that when not constrained in a laboratory setting, users (in this case, *annotators*) were more impartial. Thus, by comparing users' and annotators' datasets, we were able to detect positive bias. In future evaluations (using the same types of scenarios), we would use the annotators' data as a baseline for detecting user bias.

- *Contextual information paradox.* The negative relationship between Context_Turn and satisfaction led to an interesting discovery, in which *user skill* was identified as an important factor to this paradox. The effects of contextual turns differ substantially between good and bad speakers, whereby the former almost did not depend on the agent's contextual handling strategies (almost no correlation), as they addressed their goals in fewer turns compared to the latter. This finding suggests that Context_Turn could be a strong clue in identifying user experience.

Future work involves developing a suitable response generation model according to the various intensities of user frustration and contentment. We will also analyze the impact of the said generation model on user experience, other than user affect, by incorporating a *personality* model. For example, a novice user may prefer a dominant agent that is more verbose, explicit and directive, whilst an expert user may favour a submissive system that is more user-led, as suggested in Mairesse et al. [2007]; Reeves and Nass [1996].

Thus, the agent may need to respond differently to a *frustrated novice* user than to a *frustrated expert* one. Finally, we will conduct a series of cross evaluations between users and adaptable/non-adaptable versions of the HiFi agent and compare the findings.

10. Acknowledgement

The work leading to these results has received funding from the European Union under grant agreement num. 287678. It has also been supported by TIMPANO(TIN2011-28169-C05-03), ITALIHA(CAM-UPM), INAPRA (DPI2010-21247-C02-02), SD-TEAM (TIN2008-06856-C05-03) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. The corresponding author thanks University Science of Malaysia and the Malaysian Ministry of Higher Education for the PhD funding. Authors also thank all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics.

References

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, K., Tetreault, J., Purandare, A., 2006. Using system and user performance features to improve emotion detection in spoken tutoring systems. In: Proc. of Interspeech. pp. 797–800.

Ai, H., Raux, A., Bohus, D., Eskenazi, M., Litman, D., 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In: 8th SIGdial Workshop on Discourse and Dialogue.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. of International Conference on Spoken Language Processing (ICSLP).

Bailey, J. E., Pearson, S. W., 1983. Development of a tool for measuring and analyzing computer user satisfaction. Management Science 24, 530–545.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. Personality and Social Pscyhology 70, 614–636.

Barra-Chicote, R., Fernández, F., Lutfi, S., Lucas-Cuesta, J. M., Macias-Guarasa, J., Montero, J., San-Segundo, R., Pardo, J., September 2009. Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In: Proceedings of Interspeech. pp. 336–339.

Barra-Chicote, R., J.M. Montero, J. M., D'Haro, L., Segundo, R. S., de Cordoba, R., May 2006. Prosodic and segmental rubrics in emotion identification. In: Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1085–1088.

Barra-Chicote, R., Macias-Guarasa, J., Montero, J., Rincon, C., Fernandez, F., Cordoba, R., October 2007. In search of primary rubrics for language independent emotional speech identification. In: Proc. of WISP.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., Macias-Guarasa, J., 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Commun. 52 (5), 394 – 404.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., 2011. Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech. Comput. Speech Lang. 25 (1), 4 – 28.

Burkhardt, F., van Ballegooy, M., Engelbrecht, K.-P., Polzehl, T., Stegmann, J., 2009. Emotion detcion in dialog systems: Applications, strategies and challenges. In: Proceedings of IEEE.

Callejas, Z., López-Cózar, R., 2008a. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Commun. 50 (5), 416 – 433.

Callejas, Z., López-Cózar, R., 2008b. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems. PIT '08. Springer-Verlag, Berlin, Heidelberg, pp. 221–232.

Callejas, Z., López-Cózar, R., 2008c. Relations between de-facto criteria in the evaluation of a spoken dialogue system. Speech Commun. 50, 646–665.

Charfuelán, M., López, C. E., Gil, J. R., Rodríguez, M. C., Gómez, L. H., 2000. A general evaluation framework to assess spoken language dialog systems: Experience with call center agent systems. In: TALN.

Cowie, R., Douglas-Cowie, E., Martin, J.-C., Devillers, L., 2010. A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience and Affective Computing. Oxford University Press, Ch. The esential role of human databases for learning in and validation of affectively competent agents, pp. 151–165.

Danieli, M., Gerbino, E., 1995. Metrics for evaluating dialogue strategies in a spoken language system. In: AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation. pp. 34–39.

Devillers, L., Campbell, N., 2011. Special issue of computer speech and language on affective speech in real-life interactions. Comput. Speech Lang. 25 (1), 1 – 3, affective Speech in Real-Life Interactions.

Devillers, L., Rosset, S., Bonneau-Maynard, H., Lamel, L., 2002. Annotations for dynamic diagnosis of the dialog state. In: LREC. European Language Resources Association.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., Graesser, A., 2008. Automatic detection of learner's affect from conversational cues. User Model User-Adap. Inter 18, 45–80.

Doll, W. J., Torkzadeh, G., 1988. The measurement of end-user computing satisfaction. MIS Quarterly 12, 259–274.

Dybkjær, L., Bernsen, N. O., Minker, W., 2004. Evaluation and usability of multimodal spoken language dialogue systems. Speech Commun. 43, 33–54.

Ekman, P., Friesen, W., 1978. The Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press.

Engelbrecht, K.-P., Gödde, F., Hartard, F., Ketabdar, H., Möller, S., 2009. Modeling user satisfaction with hidden markov model. In: Proceedings of the 10th Anual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL). pp. 170–177.

Fernández-Martínez, F., Blázquez, J., Ferreiros, J., Barra-Chicote, R., Macias-Guarasa, J., Lucas-Cuesta, J. M., 2008. Evaluation of a spoken dialog system for controlling a hifi audio system. In: Proceedings of the IEEE Workshop on Spoken Language Technology. Goa, India.

Fernández-Martínez, F., Ferreiros, J., Lucas-Cuesta, J. M., Echeverry, J. D., San-Segundo, R., Córdoba, R., September 2010a. Flexible, robust and dynamic dialogue modeling with a speech dialogue interface for controlling a hi-fi audio system. In: Proceedings of the IEEE Workshop on Database and Expert Systems Applications (DEXA 2010). Springer, Bilbao, Spain.

Fernández-Martínez, F., Lucas-Cuesta, J. M., Chicote, R. B., Ferreiros, J., Macías-Guarasa, J., May 2010b. HIFI-AV: An audio-visual corpus for spoken language human-machine dialogue research in Spanish. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta.

Field, A., 2005. Discovering Statistics Using SPSS, 2nd Edition. Sage Publications.

Forbes-Riley, K., Litman, D., 2011a. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Commun. 53(9-10), 1115–1136.

Forbes-Riley, K., Litman, D., 2011b. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. Comput. Speech Lang. 25 (1), 105 – 126, affective Speech in Real-Life Interactions.

Gelbrich, K., 2009. Beyond just being dissatisfied: How angry and helpless customers react to failures when using self-service technologies. Schmalenbach Business Review 61, 40–59.

Grichkovtsova, I., Morel, M., Lacheret, A., 2012. The role of voice quality and prosodic countour in affective speech perception. Speech Commun. 54 (3), 414–429.

Grothendieck, J., Gorin, A., Borges, N., 2009. Social correlates of turn-taking behavior. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP '09. IEEE Computer Society, Washington, DC, USA, pp. 4745–4748.

Hone, K. S., Graham, R., 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). Natural Language Engineering 6(3/4), 287–305.

Kernbach, S., Schutte, N. S., 2005. The impact of service provider emotional intelligence on customer satisfaction. Journal of Services Marketing 19/7, 438–444.

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., Elenius, K., 2011. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. Comput. Speech Lang. 25 (1), 84 – 104.

Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE T. Speech. Audi. P. 13(2), 293–303.

Liscombe, J., Riccardi, G., Hakkani-Tür, D. Z., 2005. Using context to improve emotion detection in spoken dialogue systems. In: Proceedings of Interspeech. pp. 1845–1848.

Litman, D., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun. 48, 559–590.

Locke, E. A., 1976. The nature and causes of job satisfaction. Consulting Psychologists Press, Palo Alto, C.A.

Lutfi, S., Barra-Chicote, R., Lucas-Cuesta, J., Montero, J., July 2010. Nemo: Need-inspired emotional expressions within a task-independent framework. In: Proc.of Brain Inspired Cognitive Systems (BICS). Madrid, Spain.

Lutfi, S., Montero, J., Barra-Chicote, R., Lucas-Cuesta, J., Gallardo-Antolin, A., January 2009a. Expressive speech identifications based on hidden markov model. In: Proceedings of the International Conference on Health Informatics (HEALTHINF). pp. 488–494.

Lutfi, S. L., C.Sanz-Moreno, Barra-Chicote, R., Montero, J., November 2009b. Integrating a need module into a task-independent framework for modelling emotion: A theoretical approach. In: Proceedings of the Ninth International Conference on Epigenetic Robotics (EPIROB). pp. 221–222.

Mairesse, F., Walker, M. A., Mehl, M. R., Moore, R. K., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30, 457–500.

Möller, S., 2005. Quality of Telephone-Based Spoken Dialogue Systems. Springer, New York.

Möller, S., Smeele, P., Boland, H., Krebber, J., 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. Comput. Speech Lang. 21 (1), 26–53.

Nicholson, J., Takahashi, K., Nakatsu, R., 2000. Emotion recognition in speech using neural networks. Neural Comput. Appl. 9, 290–296.

Oudeyer, P. Y., 2003. The production and recoginiton of emotions in speech: features and algorithms. Int. J. Hum.-Comput. Stud. 59, 157–183.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S. A., 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics 37, 417–435.

Picard, R. W., 1999. Affective Computing for HCI. In: Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces. Vol. 1. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, pp. 829–833.

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology 88, 879–903.

Porayska-Pomsta, K., Mavrikis, M., Pain, H., 2008. Diagnosing and acting on student affect: the tutors perspective. User Model User-Adap. Inter 18 (No.1-2), 125–173.

Recommendation P.862, I., 2001. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Tech. rep., International Telecommunication Union.

Reeves, B., Nass, C., 1996. The Media Equation: How people treat computers, television and new media like real people and places. CSLI Publications, Standford.

Riccardi, G., Hakkani-Tür, D., 2005. Grounding emotions in human-machine conversational systems. In: Proceedings of 1st International Conference, Intelligent Technologies for Interactive Entertainment (INTETAIN). pp. (144–154 in Lecture Notes in Computer Science 3814 Springer, 2005).

Saris, W. E., Krosnick, J. E., Shaeffer, E. M., 2005. Comparing questions with agree/disagree response options to questions with construct-specific response options.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2012. Paralinguistics in speech and languagestate-of-the-art and the challenge. Computer Speech &amp; Language (0), –.

Shami, M., Verhelst, W., 2007. Automatic classification of expressiveness in speech: A multi-corpus study. In: Speaker Classification II. Vol. 4441 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 43–56, 10.1007/978-3-540-74122-05.

Tcherkassof, A., Bollon, T., Dubois, M., Pansu, P., Adam, J.-M., 2007. Facial expressions of emotions: a methodological contribution to the study of spontaneous and dynamic emotional faces. Journal of Social Psychology 37, 1325–1345.

Toivanen, J., Väyrynen, E., Seppänen, T., 2004. Automatic discrimination of of emotion from finnish. Lang. Speech 47, 383–412.

Truong, K. P., van Leeuwen, D. A., de Jong, F. M., 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Communication 54 (9), 1049 – 1063.

Vidrascu, L., Devillers, L., 2005. Detection of real-life emotions in call centers. In: INTERSPEECH. pp. 1841–1844.

Vogt, T., André, E., 2005. Comparing featre sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings of IEEE International Conference on Multimedia and Expo. Amsterdam, The Netherlands, pp. 474–477.

Walker, M., Kamm, A., Bol, J., 2000. Developing and testing general models of spoken dialogue system performance. In: In Proc. Language Resources and Evaluation Conference, LREC-2000.

Witten, I. H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques. Morgan-Kaufmann, San Francisco.